Behaviour & Information Technology, Vol. 25, No. 2, March-April 2006, 115 - 126

Taylor & Francis
Taylor & Francis Group

Attention web designers: You have 50 milliseconds to make a good first impression!

GITTE LINDGAARD†*, GARY FERNANDES‡, CATHY DUDEK§ and J. BROWÑ¶

Human-Oriented Technology Lab, Carleton University, Ottawa, Canada

Three studies were conducted to ascertain how quickly people form an opinion about web page visual appeal. In the first study, participants twice rated the visual appeal of web homepages presented for 500 ms each. The second study replicated the first, but participants also rated each web page on seven specific design dimensions. Visual appeal was found to be closely related to most of these. Study 3 again replicated the 500 ms condition as well as adding a 50 ms condition using the same stimuli to determine whether the first impression may be interpreted as a 'mere exposure effect' (Zajonc 1980). Throughout, visual appeal ratings were highly correlated from one phase to the next as were the correlations between the 50 ms and 500 ms conditions. Thus, visual appeal can be assessed within 50 ms, suggesting that web designers have about 50 ms to make a good first impression.

1. Introduction

First impressions have been shown to be very powerful in a wide range of contexts including studies in personality character attributions (see Anderson 1980, 1981 for numerous examples), medical diagnosis (Lindgaard and Triggs 1990, Klatzky *et al.* 1994, Eddy 1999, Ralph 2004), and studies of websites exploring perceptions of appeal and usability (Schenkman and Jönsson 2000, Tractinsky *et al.* 2000), trust (Karvonen 2000), reliability (e.g. Basso *et al.* 2001), and the relationship between several hedonic factors, beauty and what Hassenzahl (2004a) calls 'goodness'. For example, one study exploring users' experience with a website previously found to be visually extremely appealing, was equally highly valued before and after a usability test in which, on average, participants completed less than one-half of the tasks successfully (Lindgaard and Dudek 2002). Usability was clearly perceived to be very low, even before the usability test, but the strong impact of the visual appeal of the site seemed to draw attention away from usability problems. This suggests that aesthetics, or visual appeal, factors may be detected first and that these could influence how users judge subsequent experience (e.g. Jennings 2000, Tractinsky *et al.* 2000) and enjoyment with that site (van der Heijden 2003).

In the marketing research literature this long-term effect of a first impression is sometimes referred to as a 'halo effect', which carries over that first impression to the evaluation of other attributes of products (Bryant 1997). In the human decision-making and judgement literature, the phenomenon is typically referred to as a cognitive 'confirmation bias' (Mynatt *et al.* 1977, Nisbett and Ross 1980). It occurs when participants search exclusively for confirmatory evidence supporting their initial hypothesis while ignoring disconfirmatory evidence. Thus, in the presence of a very positive first impression, a person may disregard or downplay possible negative issues encountered later: potentially negative aspects such as errors may be generously overlooked (Campbell and Pisterman 1996).

Thus, in the presence of a very positive first impression, a person may disregard or downplay possible negative issues encountered later: potentially negative aspects may be generously overlooked (Campbell and Pisterman 1996). Along similar lines, a confirmation bias occurring in the context of a negative first impression will lead to failure to revise the initial hypothesis, even in the presence of strong disconfirmatory, in this case positive, evidence. Hence, even if a website is highly usable and provides very useful information presented in a logical arrangement, this may fail to impress a user whose first impression of the site was

*Corresponding author. Email: †Gitte_Lindgaard@carleton.ca; ‡gary.fernandes@rogers.com; §cathy@interactingwithcomputers.com; ¶mmjbrown@connect.carleton.ca

Behaviour & Information Technology
ISSN 0144-929X print/ISSN 1362-3001 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/01449290500330448

negative. The extent to which the strength of the first impression and a subsequent confirmation bias can be shown to generalise across different websites and across users would suggest that the impact of the feeling evoked by the first impression should not be ignored. The main objective of this paper is to ascertain whether a first impression can be formed with very brief stimulus exposure times.

1.1 *Evidence for the immediacy of responses*

Confirmation biases and the belief that the first impression is formed immediately raise the question – How immediate is immediately? Zajonc (1980) showed convincingly that stimulus preferences developed with exposure times as low as 1–5 ms (see also Moreland and Zajonc 1979, Kunst-Wilson and Zajonc 1980), and that increases in the number of exposures strengthened the effect without participants recognising previously seen stimuli. This 'mere exposure' effect has been shown to be extremely robust, occurring in several hundred studies (Bornstein 1992), and lending support to the claim that the forms of experience we call 'feeling' accompanies all cognitions. It arises early in the process of registration and retrieval (LeDoux 1996), and affective reactions that often accompany judgements of objective properties cannot be voluntarily controlled (Zajonc 1980), even though we may be able to control the expression of our feelings (LeDoux 1996). Feelings happen to us whether we like it or not, and they can apparently happen in a matter of a few milliseconds.

More recent neurophysiological evidence supports the contention that emotional responses can indeed occur pre-attentively, before the organism has had a chance cognitively to analyse or evaluate the incoming stimulus or stimuli. A small bundle of neurons has been identified that lead directly from the thalamus to the amygdala across a single synapse (Damasio 2000, p. 70; LeDoux 1992), allowing the amygdala to receive direct inputs from the sensory organs and initiate a response *before* the stimuli have been interpreted by the neocortex (LeDoux 1994).

Hence, emotions *can* apparently be triggered far more quickly than rational responses (Ekman 1992; Epstein 1994). Ekman's research on facial expression, for example, has shown that emotional expressions begin to show in changes in facial musculature within a few milliseconds after exposure to a stimulus (Ekman 1992). Even very young children exhibit fear of large, dark, noisy objects approaching rapidly on first encounter, suggesting that the potential for registering and experiencing fear is hard-wired (Barnard and Teasdale 1991; Ohman and Mineka 2001), requiring no "learning". Recognition of the object is unnecessary; all that is required is detection by the sensory system and the signaling structures – including the amygdala – to initiate some immediate response. In the absence of this autonomic body response, or in the absence of the potential to recognize the resulting body state-as 'fear', a dangerous or threatening situation would be experienced as a non-event. As LeDoux (1996) so aptly says: 'the conscious feeling we are aware of are scientifically "red herrings". Take away the subjective register of fear and there's not much to a dangerous experience' (p. 83).

It would thus appear that while rational thought makes logical connections between causes and effects, emotions are indiscriminate, connecting things that have similarly striking features (Epstein and Brodsky 1993, p. 55). Emotional "logic" is believed instead to be associative. That is, objects in the world may not necessarily be defined by their objective identity: what matters is how they are perceived. If users' perceptions, on occasion, do not reflect objective reality, then this puts further pressure on web designers to ensure that their products do create a positive first impression no matter how usable their website is and regardless of the quality of information it may contain.

1.2 *Aesthetics, beauty and visual appeal*

As noted by Lindgaard and Whitfield (2004), it is surprising that so many recent publications centring specifically on emotion in design (e.g. Green and Jordan 2002, Interactions 2004) as well as emotional theories per se, unaccountably neglect aesthetics. Some appear more or less implicitly to assume that aesthetics equates to 'beauty' or 'visual appeal' (e.g. Tractinsky *et al.* 2000, Norman 2004a); others, even integrative theories that seek to couple emotion and cognition such as affective computing, overlook it (Picard 1998).

Aesthetics, like beauty, is an elusive and confusing construct. The similarity or overlap between beauty and aesthetics remains undefined; we are unsure about what is being judged (Frohlich 2004), whether they are properties of objects in the world, subjective experiences, emotional reactions in 'the eye of the beholder', or cognitive judgements (Frolich 2004, Hassenzahl 2004a, 2004b, Norman 2004b). As Norman (2004b) points out, we sorely lack a standard body of terminology, theory and methods of investigation.

One recent paper that begins to operationalise aesthetics (Lavie and Tractinsky 2004) identifies two dimensions that the authors label 'classical' and 'expressive' aesthetics, respectively. 'Classical' aesthetics pertains to aesthetic notions dating back to antiquity and referring to orderliness in design, including concepts like 'clean', 'pleasant', 'symmetrical' and 'aesthetic'. This dimension thus contains both cognitive (clean, symmetrical) and emotional responses (pleasant). However, the fact that 'aesthetics' also appears as a dimension of aesthetics is problematic. 'Expressive' aesthetics reflects the perception of the

designers' creativity and originality, and includes concepts like 'sophisticated', 'creative', 'uses special effects' and 'fascinating'. Again, this dimension contains both types of responses. Therefore, while these concepts provide a good first step towards operationalising aesthetics, and while they may be helpful for setting explicit design goals or for assessing designs against such goals, they do not resolve the conflict of defining aesthetics clearly and explicitly. Since the purpose of this paper is to determine the immediacy of a first impression rather than to define aesthetics, the term 'visual appeal' is used here to denote what many would call 'aesthetics', and which may consist of both 'classical' and 'expressive' components. The problem of defining aesthetics is not addressed further.

1.3 *The appraisal of visual appeal*

In addition to Lavie and Tractinsky's (2004) work, other authors also argue that the appraisal of visual appeal comprises several dimensions. For example, Creusen and Snelders (2002) developed a set of three scales taking this into account. One, they refer to as the 'hedonic' scale, which measures emotion-related aspects of buying decisions; another, dealing with the logic of buying decisions, is called the 'rational' scale; and the third, the 'general involvement' scale, contains items about the importance of the product and the time and effort involved in buying it. In an earlier study, Snelders (1995, cited in Creusen and Snelders 2002) found that the hedonic and the rational scales were not correlated, but that both correlated with the general involvement scale, suggesting that 'consumers think of pleasure as a separate product value, unrelated to objective product functions, but just as important to them [*as the objective (cognitive) value*]' (p. 70, italics added). Pleasure, they conclude, is not simply the end result of rational deliberation: consumers apply both holistic (emotional) and analytic (cognitive) judgement in the decision to buy a product.

A recent study (Hassenzahl 2004a) investigated the interplay between two evaluative constructs, namely beauty and 'goodness', and three sets of hedonic attributes: identification, stimulation and pragmatic quality. Using his earlier developed semantic differential scales and stimuli comprising a set of MP3-player skins, Hassenzahl found that beauty as an evaluative construct was predominantly related to a product's ability to provide identification. Identification attributes are primarily social, including judgements of perceived appearance (professional, classy, valuable and so on). Note the similarity of Hassenzahl's concept of beauty to Creusen and Snelders 'hedonic' measure. By contrast, 'goodness', which comprises perceived usability and mental effort, appeared to be more closely related to pragmatic hedonic attributes, especially when participants were also required to interact with the stimuli. His findings can be seen to agree with Creusen and Snelders' division between emotional/holistic and more considered cognitive responses. Hassenzahl argues that initial judgements of beauty without interactive experience are likely to be based on diffuse (emotional) hedonic identification attributes whereas hedonic pragmatic attributes are judged on experience-based (cognitive) quality judgements.

Creusen and Snelders' (2002) and, to some extent, Hassenzahl's (2004a) claims, echo earlier findings reported by Pickford (1972, cited in Lavie and Tractinsky 2004) in which the author proposed three levels of evaluation in the development of aesthetic preferences: an emotional level, a perceptual level and an 'aesthetic' level, which is an integration of the two first levels. Some aspects of Pickford's classification, Zajonc's (1980) early results, and Creusen and Snelders' (2002) findings share similarities with Norman's (2004a) discussion of emotional design in which he distinguishes between visceral, behavioural and reflective responses. Norman's (2004a, 2004b) behavioural responses rely both on pleasure and effectiveness of use, with 'effectiveness' corresponding to standard usability criteria, and to Creusen and Snelder's 'rational' scale. Norman's reflective response, considering the 'rationalisation and intellectualisation of a product' (2004a, p. 5), appears to be captured in Creusen and Snelder's 'general involvement' scale, and in Pickford's 'aesthetic' evaluation level. This also corresponds with Hassenzahl's idea that judgements of beauty may evolve from representing the immediate impression of appearance to an expression of the pleasure of interacting with a product.

In Norman's model, the visceral response is immediate, holistic and physiological. It is the emotional, perhaps 'mere exposure effect' (Zajonc 1980), arousal-based response that Berlyne (1971, 1972) referred to in his early work on experimental aesthetics, and possibly captured in Creusen and Snelder's 'hedonic' scale. While Hassenzahl (2004b) rejects the existence of what Norman calls 'visceral beauty', he nevertheless agrees that 'initial reactions of liking and disliking are apparent', but does not think that 'we can call these reactions beauty' (p. 381). Thus, the confusion seems to lie in the differential use of terminology rather than in the substance of the various arguments. The discussion points to agreement that the first impression is physiological, reflecting 'what my body tells me to feel' rather than 'what my brain tells me to think', with cognitive appraisal occurring after this first physiological response.

In an effort to identify some general characteristics that may affect the immediate impression of visual appeal, this study exposed participants to a large number of website homepages. In line with Tractinsky *et al.* (2000), it also aimed to determine the reliability of judgements of visual

appeal. Following Zajonc (1980) and Bornstein's (1992) findings, exposure times in the first study were long enough for participants to form a first impression. We also believed, at the time, that they were short enough to ensure appeal ratings would be relatively uncontaminated by impressions unrelated to visual appeal such as the semantic content of web page text.

2. Method

2.1 *Overview*

In the first two studies, participants viewed website homepages sequentially for 500 ms each and rated the visual appeal of each page. In the first study, 100 homepages, collected purely as best and worst examples of visually appealing web pages by members of the Human Oriented Technology Lab (HOT Lab), were presented in different random orders for each participant and every phase. Participants viewed and rated every homepage in two phases to check the consistency of the ratings. In Study 2, a group of different participants followed the same procedure to view the 25 highest-rated and 25 lowest-rated homepages as determined by Study 1, presented in different random orders for each participant and for every phase. After each participant had viewed every web page twice in Study 2, they viewed each page a third time, but this time for as long as they liked. While viewing each page, they assigned ratings to seven visual design characteristics. The purpose of the first study was to determine the reliability of visual appeal ratings and select a subset of website homepages to use in the second study. The second study had two purposes – to determine the reliability of visual appeal ratings of the subset of 50 homepages and to begin to explore visual characteristics that may be related to visual appeal.

2.2 *Rating scales and opinions of design features*

Opinions in behavioural science are typically expressed in a numeric form such as a number along a 5-point, 7-point or 10-point Likert scale, or on an interval scale usually ranging from 0 to 100. Estimates along the 5- and 7-point scales have been shown to be nonlinear (Virtanen *et al.* 1995). That is the psychological distance between, say, a rating of '2' and '3' may thus not resemble the psychological distance between a rating of '4' and '5'. In addition, the strong tendency of participants to favour the centre and avoid the extremes of scales thereby exhibiting 'conservatism' (Edwards 1999) reduces the apparent 'discriminability' and hence the possibility of finding significant differences between stimuli even if subjectively they do differ substantially. Furthermore, in the subjective probability literature it is claimed that the requirement to provide a number may not accurately reflect the participant's opinion (Fischhoff and Bruine de Bruin 1999; Bruine de Bruin *et al.* 2000). For these reasons, the validated technique of providing an unmarked line (Levin 1975, 1976, Lockhead 1992) anchored at each end by appropriate expressions by 'very unattractive' and 'very attractive' was used to collect opinions instead of conventional rating scales. The studies did not involve subjective probabilities, which are typically used when researchers are interested in an absolute judgement. Rather, we were interested in using a scale that would reveal the relationships between homepages. In Study 3, which replicated parts of study 2, we used a 9-point rating scale to ascertain whether the relationships we observed using an unmarked line would emerge as clearly.

3. Study 1

3.1 *Participants*

Participants were 22 university students who reported that they were not colour-blind and had normal vision, after correction in some cases. To participate in the study, participants spoke English as their first language. Approval for conducting this research with human participants was granted by the Ethics Committee, Department of Psychology, Carleton University.

3.2 *Apparatus*

Each participant was tested on a workstation with 1.6GHz Athlon CPU, 256 Mbytes of RAM, Matrox dual-head video card, and a Samsung SyncMaster 950p 19-inch monitor with a white balance calibrated at 9300° K and a gamma value of 2.1. A program created in Microsoft Visual Basic 6.0 was used to present images of website homepages and to collect ratings.

3.3 *Materials*

The stimuli were screen shots of 100 web pages that would not have received wide public exposure and that varied in visual appeal. The stimulus web pages were selected from a number of sources. Members of the Human Oriented Technology Lab (HOT Lab) were asked to submit links to web pages that they thought 'looked really good or looked really bad' and that they did not think to be high traffic sites (e.g. not cnn.com, amazon.com, etc.). Other web pages were brought to our attention through email distribution lists such as UK-usability. Web pages came from a variety of contexts including entertainment, e-commerce, information, personal sites, etc. Screen shots of each web page were taken within an Internet Explorer 6 browser at 1024 × 768 pixel resolution in 32-bit true colour. In the Visual Basic program, the web page images looked like they were being viewed in the Internet Explorer browser.

3.4 *Procedure*

Participants were tested individually in sessions lasting approximately 30 minutes. After reading a briefing form and signing an informed consent sheet, each participant was seated in front of a computer. They adjusted their seating height and monitor angle to their preference. Each participant saw the first stimulus web page for 500 ms. followed by a white screen with the continuous rating scale shown in Figure 1.

Participants were instructed to 'Rate the visual appeal of the web page' by clicking on the bar at the appropriate location to indicate their rating. The scale is shown in Figure 1 as it first appeared with the marker at the centre. The location clicked was recorded as a number from 0 to 100 by the Visual Basic program. The program advanced to the next stimulus web page only after the participant had entered a response followed by a 1000 ms. delay. This procedure was followed for all phases.

There were two test phases. Each test session began with 20 practice phases intended to accustom participants to the task. The practice phases used the same 20 web page images presented in a fixed order for each participant. In the first test phase, each participant then viewed 100 test phases presented in random order. In the second test phase, each participant viewed the 100 web pages for a second time in a newly randomised order. The second phase served as a reliability check.

3.5 *Results*

To check the reliability of participants' responses to the same web pages presented in the two test phases, correlations were first calculated for each participant's score on the first and the second phase. As can be seen in Table 1, one half of the correlations fell between $r = .80$ and $r = .89$, with only four participants' (18.19%) correlations falling below $r = .70$, and none falling below $r = .50$. Without exception, all correlations as well as the squared correlations were highly significant at the $p < .001$ level. Thus, participants' ratings were highly reliable.

As recommended by Monk (2004), the data were also analysed by stimuli. Accordingly, Figure 2 shows the relation between mean visual appeal ratings for each web page collapsed across all 22 participants for the first test phase and the second test phase. Data points thus are the 100 web pages. The squared Pearson Product Moment correlation coefficient (r) was .97 ($p < .001$), indicating that 94% of the variance in visual appeal ratings for the same web pages in one phase was shared with visual appeal ratings in the other phase.

4. Study 2

4.1 *Participants*

Participants were 31 students of a similar description to those in Study 1, who had not participated in Study 1.

Table 1. Correlations between each participant's score in test phases 1 and 2.

Correlation	N (and %) participants
.50–.59	1 (4.55%)***
.60–.69	3 (13.64%)***
.70–.79	6 (27.27%)***
.80–.89	11 (50%)***
.90–.99	1 (4.50%)***

***$p < .001$.

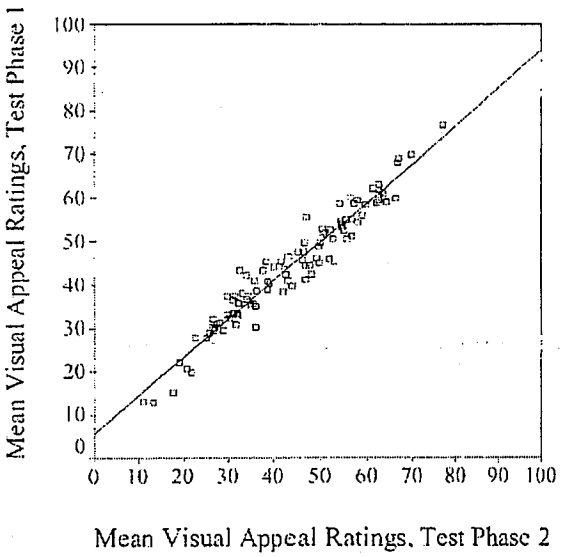

Figure 2. Relation between visual appeal ratings in the two phases (mean first rating and mean second rating for each of the 100 web pages in Study 1) ($r^2 = .94$).

Very Unattractive ——— Very Attractive

Figure 1. Visual appeal scale – participants clicked on the slider bar to make their ratings.

4.2 *Apparatus*

The same computer system and modified software from Study 1 were used except that a second computer monitor was provided and a third phase of judgements was added.

4.3 *Materials*

A subset of 50 web pages from Study 1 was used in Study 2. For each participant in Study 1, visual appeal responses for web pages within each phase were transformed into z-scores. The mean of each web page's z-scores was calculated across the two phases for each participant. This mean provided a measure of visual appeal for each homepage for each participant. Then, the medians of these means were calculated for each homepage across participants. These medians provided a general visual appeal score for each homepage. Next, web pages were ranked by their medians. The 25 least appealing and 25 most appealing web pages were selected for use in Study 2. In addition to being ranked at the top or bottom in terms of visual appeal, the 25 most appealing homepages had to fall in the top 50 for at least half of the participants, and the 25 least appealing homepages had to fall in the bottom 50 for at least half of the participants.

4.4 *Procedure*

The procedure for Study 2 was identical to Study 1 except that a test phase requiring participants to judge seven design characteristics other than visual appeal was added. After completing the first two phases as before, participants viewed each page individually again for as long as they wished while offering their opinion on each of the seven design characteristics (simple – complex; interesting – boring; clear – confusing; well designed – poorly designed; good use of colour – bad use of colour; good layout – bad layout; imaginative – unimaginative). These terms were presented in the continuous scale as in Study 1. Each website image was shown on the left monitor while the rating scales were presented on a smaller monitor to the right. After making a rating on each design characteristic, the participant pressed a 'Next' button to advance to the next web page. Since all judgements were captured electronically, there was no risk of data transcription errors. A session lasted approximately 50 minutes.

4.5 *Results*

4.5.1 Reliability of visual appeal ratings. As in Study 1, correlations were first calculated for each participant's score on the first and the second phase to check the reliability of within-subject responses in the two test phases. Table 2 shows a similar distribution as in Study 1: nearly one-half of the correlations fell between $r = .80$ and $r = .89$, and only three falling below $r = .70$, but with nine, or nearly 30% being higher than $r = .90$. Every one of the correlations as well as the squared correlations was significant at the .001 level. As before, it can safely be concluded that participants' ratings were highly reliable.

As before, the reliability of the participants' mean visual appeal ratings for the same web pages in phase 1 and phase 2 was assessed next. Figure 3 shows the relation between visual appeal ratings for the two phases based on mean ratings of 31 participants. The squared correlation, $r^2 = .97$, $p < .001$, was comparable to that obtained in Study 1. Figure 4 shows the absence of visual appeal ratings in the middle of the scale. As in Study 1, the 25 most appealing homepages were also the 25 most appealing in Study 2 and the same was true for the 25 least appealing homepages. Likewise, none of the homepages originally found to be

Table 2. Correlations between each participant's score in test phases 1 and 2.

Correlation	N (and %) participants
.50–.59	2 (6.45%)***
.60–.69	1 (3.22%)***
.70–.79	4 (12.90%)***
.80–.89	15 (48.39%)***
.90–.99	9 (29.03%)***

***$p < .001$.

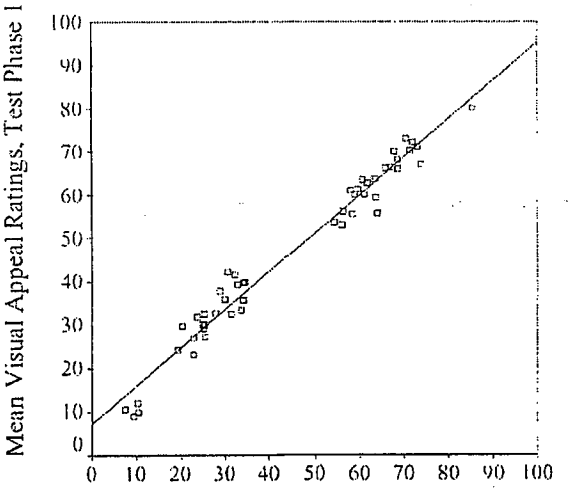

Figure 3. Relation between visual appeal ratings in Test Phases 1 and 2 (mean first rating and mean second rating for each of the 50 web pages – 25 appealing and 25 unappealing – in Study 2).

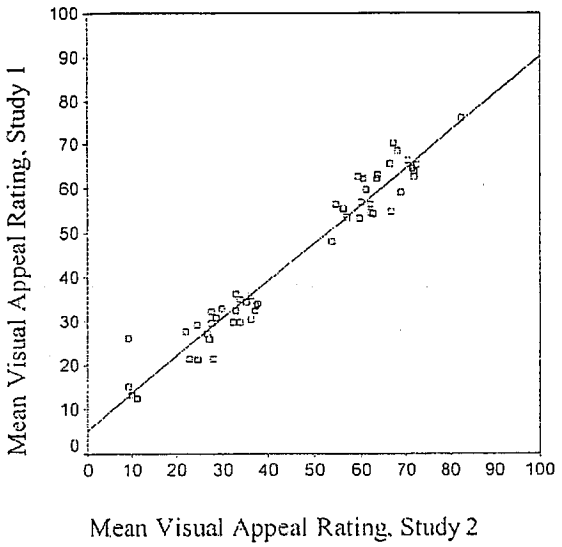

Figure 4. Relation between visual appeal ratings in Study 1 (mean of Test Phases 1 and 2) and Study 2 (mean of Test Phases 1 and 2).

unappealing were found to be appealing in Study 2, and the ranking of homepages was the same in the two studies.

As an additional check to explore if very small samples of subjects would yield equally reliable results, correlations were computed for several samples of scores obtained from five randomly selected participants. All of these were found to be equally reliable as the complete sample.

Correlations for mean visual appeal ratings for the 50 homepages in common between Study 1 and Study 2 were calculated by collapsing across Test Phases 1 and 2 for both studies. Figure 4 shows a strong relation between visual appeal ratings in the first and the second study, $r^2 = .95$, $p < .001$.

4.5.2 Relation between visual appeal ratings and other design characteristics. The first part of Study 2 provided highly reliable ratings of mean visual appeal for each of 50 web pages. The second part also provided mean ratings on seven design characteristics for each web page in an effort to begin to uncover the relationship between these and visual appeal. Seven zero-order Pearson Product Moment squared correlations were calculated with the 50 web pages as cases. For each squared correlation, one variable was the mean homepage visual appeal rating by each of the 31 participants collapsed across all 50 homepages and the other the mean rating by the same 31 participants for a visual characteristic. There were very high squared correlations between visual appeal and five of the seven visual characteristics: interesting–boring ($r^2 = .91$, $p < .001$), good design–bad design ($r^2 = .92$, $p < .001$), good colour–bad colour ($r^2 = .90$, $p < .001$), good layout–bad layout ($r^2 = .88$, $p < .001$), and imaginative–unimaginative ($r^2 = .86$, $p < .001$). There was a low correlation between visual appeal and simple–complex ($r^2 = .01$, $p > .80$) and a moderate correlation between attractiveness and clear–confusing judgements ($r^2 = .39$, $p < .001$).

The five visual characteristics that were very highly correlated with visual appeal were also highly correlated with each other with squared correlations ranging from $r^2 = .82$ to $r^2 = .97$. A multiple regression was performed predicting mean visual appeal from ratings on interesting–boring, good design–bad design, good colour–bad colour, good layout–bad layout, and imaginative–unimaginative. The best linear combination of these characteristics was very highly correlated with rated visual appeal, $R^2 = .94$, $p < .001$.

4.5.3 Graphical properties determining visual appeal. An attempt was made to identify the graphical properties that underlie judgments of visual appeal and to demonstrate the validity of these judgements by means of comparing them to expert designers' judgements. Two experts evaluated aspects of each of the categories. Eighty-nine properties were compared. Of these, only nine had interrater correlations above $r = .70$ meaning that, in general, the experts did not agree on the properties of the web pages.

Of the nine properties on which the experts did agree, only five had sufficient variability in the sample of homepages. For example, the property 'screen dominance' had sufficient variability because some homepages had a very good balance between text and graphics, others had very bad balance and yet others were somewhere in-between. For five properties, an analysis could be conducted to see if the experts' scores on these tended to result in high or low visual appeal scores. Regressing visual appeal on the five properties, Fernandes (2003) concluded that a combination of these properties determines visual appeal as opposed to any one of the properties on its own. However, of the five properties, 'screen dominance' was most significant. A measure of 'how carefully and discreetly display techniques are used together' was also significant, but this really just supported the overall finding that a combination of factors predicted visual appeal. It was impossible to validate this measure.

4.6 *Discussion*

4.6.1 Visual appeal, reliability of judgements and individual differences. It is often said that 'beauty is in the eye of the beholder' with large individual differences in what people like and don't like. Indeed, as Hassenzahl (2004b) points out, two people may find the same object beautiful or ugly for the same reason, perhaps because the object fits one but not the other person's individual style. It is therefore

tempting to assume that perception of website visual appeal would result in as many different opinions as there are people. The above results suggest that a relatively small number of people in aggregate can reach remarkable agreement on the visual appeal of homepages. Large correlations are rare in much of the social sciences and in participative judgements in general. Squared correlations of .90 or higher are rare – yet we consistently found high reliability of mean appeal judgements with squared correlations ranging from $r^2 = .89$ to $r^2 = .95$.

4.6.2 Relation between visual appeal and other design characteristics. Our results suggest that there appears to be a strong relationship between visual appeal and several other design characteristics. However, our selection of design characteristics was simply based on previous research in our lab (Tombaugh *et al.* 1982). While the findings are interesting, the relationships among design elements and visual appeal deserve a much more systematic and careful analysis of possible design characteristics. As with the notions of aesthetics and beauty, we are confronting ambiguity in terminology defining such characteristics. For example, our interesting – boring scale may be what Hassenzahl (2004a) refers to in his 'lame – exciting' scale; our imaginative – unimaginative continuum could be either 'amateurish – professional' or 'standard creative' in Hassenzahl's language. and our good design – bad design may capture some, but certainly not all of Hassenzahl's concept of 'goodness'. In Hassenzahl's research, these concepts belong in different hedonic quality categories. which he showed to impact differently on beauty and goodness. However, we need more precise definitions as well as quantification of the concepts that contribute to judgements of design characteristics including beauty and goodness. An excellent and promising start has been made to identify quantitative relationships between key design characteristics and generic dimensions of emotions typically experienced when inspecting homepages (Kim *et al.* 2003). From their list of emotions, they identified a set of key design factors that professional designers use when developing emotionally evocative homepages. Bringing the two together enabled the researchers to quantify relationships between them and use these to analyse homepages developed by their team of professional designers. Using the Kim *et al.* recommendations, we are now attempting to analyse the homepages used in the above studies.

The attempt to obtain information from expert designers on the individual design features that may have influenced participants' ratings was unsuccessful. Originally, it was intended that a group of designers working independently would assess the design features present and/or absent from each website in order to isolate specific design features that would seem to affect participants' judgements of visual appeal. However, the task of doing this was extremely time-consuming, and it quickly became obvious that the eight designers who completed part of that analysis disagreed so vehemently that it would be impossible to identify specific 'principles' that were either systematically taken into account or violated in each of the 50 homepages.

The above studies suggest that 500 ms was short enough to form a first impression, but possibly also long enough to allow cognitive processing of specific attributes such as ease-of-use and purpose. Anecdotal observations suggested that at least some participants 'saw' details that went beyond the first holistic response. Indeed, the studies reviewed by Bornstein (1992) led him to conclude that the mere exposure effect begins to wane at 50-ms stimulus exposure times. Study 3 was therefore designed to test whether the first impression may be formed in an even shorter time than 500 ms.

5. Study 3

Using the same stimuli as in Study 2, participants in Study 3 saw the homepages for 50 or 500 ms. A sample of 40 participants was randomly assigned to one condition only, either 50 ms ($n = 20$) or 500 ms ($n = 20$). The 500-ms exposure time was included to allow comparison of the two conditions as a between-subject variable. Since Study 2 clearly demonstrated a relationship between five of the seven attributes tested and visual appeal, and because further research is needed to understand the role of the remaining two attributes, only visual appeal was rated in Study 3. As in Study 1, the 20 practice pages were shown first. Thereafter, each homepage was again shown and rated in two exposures, presented in a different random order for each of the 40 participants none of whom had taken part in any of the previous studies.

5.1 *Apparatus*

Each participant was tested on a workstation with 1.1 GHz Athlon CPU, 512 Mbytes of RAM, RADEON 7000 Series video card, and a ViewSonic Graphics Series G90f 19-inch monitor with a white balance calibrated at 9300° K and a gamma value of 2.1 with resolution set to 1024 by 768 pixels. A program created in DirectRT™ was used to present images of website homepages and collect ratings.

5.2 *Procedure*

Participants were tested individually in sessions lasting approximately 25 minutes. The procedure was exactly the same as in Study 1 with the exception that, instead of the unmarked line, participants responded using numeric keys 1–9 on the keyboard, where they were told in the instructions that 1 = 'very unappealing' and 9 = 'very

appealing'. After each homepage was displayed a screen with the words 'rate appeal (Use keys 1 through 9)' was displayed, and at that point the participant pressed the key that best represented their opinion. After the key was depressed a blank screen was shown for 1000 ms, and then the next homepage was displayed.

5.3 *Results*

Overall, the results appeared to resemble those obtained in Study 2. In order to address the crucial research question as to whether a first impression of homepages can be formed in less than 500 ms, a Pearson Product Moment correlation comparing the mean visual appeal ratings on the first phase for the two conditions (50 and 500 ms) was calculated. Scores were collapsed across all homepages. It revealed that $r = .947$, $p < .001$ ($r^2 = .897$). This result was slightly lower, but comparable to that obtained in Study 2 for the 500 ms condition alone. Then re-analysing the data, using the median instead of the mean rating, resulted in $r = .911$ ($r^2 = .83$) for the first phase of both conditions. Likewise, the correlation for the second phase of both conditions yielded a value of $r = .953$, $p < .001$, ($r^2 = .908$) and again, using the median instead of the mean ratings, resulted in $r = .922$ ($r^2 = .85$). Findings would thus appear to be as robust with an exposure time of 50 ms as with an exposure time of 500 ms.

A more detailed analysis of the data was performed comparing the 50-ms and the 500-ms conditions. A correlation of the interrater reliability compared each participant's ratings of the 50 homepages with each of the other 19 participants in each of the two phases. The average correlation for each phase was computed yielding $r = .557$ at 500 ms on the first phase, and $r = .599$ on the second phase. In the 50-ms condition, the average correlation was $r = .337$ on the first phase and $r = .403$ on the second. In both cases, the correlations thus increased between first and second rating.

For each participant, the number of insignificant correlations was counted to determine the extent to which each participant agreed with the 19 others. Table 3 suggests that the percentage of insignificant correlations was higher in the 50-ms condition than in the 500-ms condition for both phases 1 and 2. Hence, the variability among participants was substantially greater in the 50-ms than in the 500-ms condition, and overall, participants were considerably more consistent from one phase to the next in the 500-ms than in the 50-ms condition. To deal with the theoretical properties of correlations of distributions, the correlations were transformed using the formula $z = 1/2 \ln (1 + r) - 1/2 \ln (1 - r)$ suggested by MacNemar (1969, p 147). Raw correlations between phase 1 and phase 2 increased for the 50-ms case ($M = 0.066$, $SD = 0.069$) and for the 500-ms case ($M = 0.042$, $SD = 0.053$). In both cases this increase was statistically significant, $t(19) = 4.461$, $p < .001$, two-tailed and $t(19) = -3.606$, $p < .01$, two-tailed, respectively.

As in Studies 1 and 2, participant's scores for the first and second test phase were correlated for each of the 50-ms and 500-ms conditions as is shown in Table 4.

There was a clear difference in the distributions of the two conditions; in the 50-ms condition 40% of the correlations were below $r = .60$ whereas none was below $r = .60$ in the 500-ms condition. Yet even so, some 60% of the correlations were between $r = .60$ and $r = .79$ in the 50-ms condition, and, as in Studies 1 and 2, the bulk of correlations were above $r = .80$ in the 500-ms condition. Despite this spread, all correlations but one were significant in both conditions; in the 500-ms condition, all were significant at the $p < .001$ level. In the 50-ms condition, all but three were significant at the $p < .001$ level; two were significant at the $p < .05$, and one was not significant.

6. General discussion

6.1 *First impressions = mere exposure effects?*

The above findings demonstrate that participants reliably decided which homepages they liked and which ones they did not like within 50 ms as evidenced by the highly significant correlations between phases 1 and 2 in both the 50-ms and all the 500-ms conditions. First impressions of

Table 3. Percentage and raw counts of insignificant correlations.

Condition	Phase 1	Phase 2
50 ms	41.0% (148)	28.8% (104)
500 ms	2.8% (10)	2.8% (10)

Table 4. Number and percentage of participants with significant correlations between phases 1 and 2.

Correlation	N participants and (%) 50-ms condition	N participants and (%) 500-ms condition
.10–.19	1 (5%)	
.20–.29	2 (10%)*	
.30–.39	2 (10%)***	
.40–.49	1 (5%)***	
.50–.59	2 (10%)***	
.60–.69	8 (40%)***	1 (5%)***
.70–.79	4 (20%)***	4 (20%)***
.80–.89		15 (75%)***

*$p < .05$; ***$p < .001$.

homepages would thus seem to be formed in a time-frame that Bornstein (1992) and Zajonc (1980) would regard as a mere exposure effect, representing a holistic, physiological (LeDoux 1996, Damasio 2000) response. The similarity in ratings between the three studies at 500 ms as well as between first and second ratings in all of these studies testifies to the robustness of the findings at least for the sample of homepages tested here. The fact that judgements between participants were more consistent at the 500-ms level than at the 50-ms level, may be due to do with a differential amount of information perceived in the two conditions; it is possible that at 500 ms participants were taking in much more information related to content and purpose of the page than was true in the 50-ms condition. This reasoning could help to explain another interesting finding – the fact that in both the 50-ms condition and the 500-ms condition, the level of agreement between participants appeared to increase substantially from the first to the second phase.

It is possible that the mere exposure effect begins to wane even before a stimulus has been viewed for 50 ms, enabling participants to attend to more design characteristics with every exposure. We do not intend to pursue this argument further; our aim is not to determine an accurate threshold of first impressions, but to ascertain whether a first impression can be formed reliably in less than 500-ms exposure times and thus constitute a mere exposure effect. The findings appear to support both of these goals. For that same reason, the homepages falling between the two extremes in Study 1 were eliminated in Studies 2 and 3: we were not interested in determining the reliability of judgements of homepages falling between the very appealing and very unappealing homepages.

6.2 *Is there a visceral beauty?*

Norman (2004b) asserts that 'at the visceral level, there can only be positive and negative valence and these can only be assessed through physiological measurements. Any spoken or conscious assessment of visceral responses must come from the reflective level, which means it has been subjected to possible interpretation, modification and rationalisation' (p. 315). Rating the visual appeal of a homepage is indeed a considered response of sorts, but is it really possible to modify and rationalise one's impression of a stimulus, literally seen at a glimpse during which one cannot possibly discern all its details? On the one hand, our results support the notion that participants did more than merely decide whether each homepage evoked a diffuse 'good' or 'bad' feeling. Even that level of interpretation would go beyond Norman's strict definition of visceral beauty, as participants were required to register their impression and place a judgement on a scale. Had the judgements been an all-or-none response, one would have expected half the judgements to be tightly clustered around the very low end, and the other half around the extreme high end of the scale. That clearly did not happen. It is, of course, quite possible that individuals are internally consistent, producing very similar judgements on the same stimuli in two phases, and that the spread of scores simply represents individual differences in the way participants used the scales. On the other hand, it is also possible that individuals make relatively 'uninformed' judgements on the basis of a minimum of information, without engaging in any form of deep cognitive and conscious reflection. Is it not possible that participants were employing what Damasio (2000) calls 'somatic markers' – emotional thermometers by which he claims we assess our immediate emotional responses to situations or stimuli enabling us to deal with these with a minimum of cognitive energy? Maybe we are so accustomed to applying such somatic markers that they reliably tell us 'how good' or 'how bad' our response to a given stimulus feels, and maybe we rely on these in situations where there is not time consciously to scrutinise the perceived stimulus.

Hassenzahl (2004b) dismisses the possible existence of a visceral beauty, arguing that the kind of valenced, affective response that Zajonc (1980) showed could be made without cognitive involvement, may not represent a complex emotion like hate or love. However, Norman makes no claims about the complexity of the diffuse subconscious-level emotion evoked viscerally, saying instead that 'it is only at the reflective level that full-fledged emotions reside' (2004b, p. 315) and that this level is conscious, intellectually driven and aware of emotional feelings. We are not convinced that our results support this last statement. Both researchers agree that beauty judgements are interpretations of 'initial, diffuse, spontaneous responses of liking and disliking' (Hassenzahl 2004b, p. 381). Agreeing, as Norman goes on to suggest, to restrict the term beauty to conscious, reflective judgements may bring us a little closer to a crisper definition and settle some of the ambiguity inherent in the term, but how are we to interpret our results? Perhaps the next steps should involve alternative procedures such as eye tracking or taking physiological measures. Clearly, research into this interplay between emotion and cognition is in its infancy.

6.3 *The rating scale saga*

As discussed earlier, the requirement to express subjective probabilities as a number representing one's opinion can be problematic. Likewise, expressing an opinion using numbered rating scales may fail to represent participants' true opinions because such scales have been shown to be psychologically nonlinear. We argued that, since the issue here was to learn whether the relationship between the sample of homepages used in the above studies would be

similar when using a rating scale or a continuous line, the rating scale was used in Study 3. The within-subject reliability had already been demonstrated for this sample of homepages using the continuous line in the previous studies. Study 3 suggested that the relationship was very similar across all experiments for the 500-ms condition. Because there were several changes, albeit all relatively minor, between Studies 1, 2 and 3, it was not possible to verify this relationship statistically. However, all results clearly showed that participants' judgements on the two phases were highly reliable; once they had decided how much they liked what they saw, they tended to stick with that same judgement in the next exposure. The above data thus provide some evidence suggesting that it is safe to use rating scales in situations in which relationships rather than absolute judgements are investigated and in which intrarater reliability is at issue.

7. Conclusion

Our ambition was to determine how quickly people decide whether they like or dislike what they see, and whether such judgements may constitute a mere exposure effect. The above data suggest that a reliable decision can be made in 50 ms, which supports the contention that judgements of visual appeal could represent a mere exposure effect. The level of agreement between participants and between experiments was impressive and highly correlated even for the 50-ms condition. Our data also suggest that the notion of visual appeal may be closely related to other concepts concerning overall impressions of design layout, colour and so forth. However, more research is needed to establish the nature of these relationships more accurately.

Our second ambition was to begin to understand what specific design attributes may contribute to visual appeal. That was too hard to do, at least using the method we employed here, and probably because the relationship between individual design features and the first holistic impression may not be as simple as we thought. Instead, we are now re-analysing the data using the Kim *et al.* (2003) technique.

It is clear from these studies that first impressions form quickly and are consistent. The strength of the results presented here suggests that designers should be very interested in finding out what, if any, effect the immediate first impression has on subsequent behaviours, such as selecting a site or buying from one. The question that should be resonating in the minds of all web designers is – how much weight does this first impression carry? Clearly, more research is needed to address that question.

Acknowledgement

We would like to thank the editors and the two anonymous reviewers for their helpful, constructive comments.

References

ANDERSON, N.H., 1980, *Foundations of information integration theory* (Sydney, Academic Press), pp. 85–94.

ANDERSON, N.H., 1981, *Foundations of information integration theory* (London: Academic Press).

BARNARD, P.J. and TEASDALE, J.D., 1991, Interacting cognitive subsystems: A systematic approach to cognitive-affective interaction and change. *Cognition and Memory*, **5**, pp. 1–39.

BASSO, A., GOLDBERG, D., GREENSPAN, S. and WEIMER, D., 2001, First impressions: Emotional and cognitive factors underlying judgments of trust in e-commerce. *Proceedings of the 3rd ACM conference on Electronic Commerce* (New York: ACM Press), pp. 147–143.

BERLYNE, D.E., 1971, *Aesthetics and Psychobiology* (New York: Appleton-Century-Crofts), Chapter 3.

BERLYNE, D.E., 1972, Experimental aesthetics. In *New Horizons in Psychology*, P.C. Dodwell (Ed.), pp. 9–22 (Harmondsworth: Penguin).

BORNSTEIN, R.F., 1992, Subliminal Mere Exposure Effects. In *Perception Without Awareness: Cognitive, Clinical, and Social Perspectives*, R.F. Bornstein and T.S. Pittman (Eds.), pp. 191–210 (New York: The Guilford Press).

BRUINE DE BRUIN, W., FISCHHOFF, B., MILLSTEIN, S.G. and HALPERN-FELSHER, B.L., 2000, Verbal and numerical expressions of probability: "It's a fifty-fifty chance". *Organisational behaviour and human decision processes*, **81**, pp. 115–131.

CAMPBELL, A. and PISTERMAN, S., 1996, A Fitting Approach to Interactive Service Design, The Importance of Emotional Needs. *Design Management Journal*, **Fall**, pp. 10 - 14.

CREUSEN, M. and SNELDERS, S., 2002, Product appearance and consumer pleasure. In *Pleasure with Products: Beyond Usability*, W.D. Green and P.W. Jordon (Eds.), pp. 69–75 (New York: Taylor and Francis).

DAMASIO, A.R., 2000, A second chance for emotion. In *Cognitive Neuroscience of Emotions*, R.D. Lane and L. Nadel (Eds.), pp. 12–24 (New York: Oxford University Press).

EDDY, D.M., 1999, Probabilistic reasoning in clinical medicine: Problems and opportunities. In *Judgment under uncertainty: Heuristics and biases*, D. Kahneman, P. Slovic and A. Tversky (Eds.), pp. 249–268 (Cambridge, UK: Cambridge University Press).

EDWARDS, W., 1999, Conservatism in human information processing. In *Judgment under uncertainty: Heuristics and biases*, D. Kahneman, P. Slovic and A. Tversky (Eds.), (Cambridge, UK: Cambridge University Press).

EKMAN, P., 1992, An argument for basic emotions. *Cognition and Emotion*, 6, pp. 169–200.

EPSTEIN, S., 1994, Integration of the cognitive and psychodynamic. *American Psychologist*, **44**, pp. 709–724.

EPSTEIN, S. and BRODSKY, A., 1993, *You're smarter than you think* (New York: Simon & Schuster), pp. 38–46.

FERNANDES, G.J., 2003, Judging web page visual appeal, unpublished Masters Thesis, Carleton University, Ottawa, Canada.

FISCHHOFF, B. and BRUINE DE BRUIN, W., 1999, Fifty-fifty = 50%? *Journal of Behavioral Decision Making*, **12**, pp. 149–163.

FROHLICH, D.M., 2004, Beauty as a design prize. *Human–Computer interaction*, **19**, pp. 359–366.

Green, W.S. and Jordan, P.W. (Eds.), 2002, *Pleasure with products: Beyond usability* (New York: Taylor & Francis), pp. 1–9.

HASSENZAHL, M, 2004a, The interplay of beauty, goodness, and usability in interactive products. *Human–Computer interaction*, **19**, pp. 319–349.

HASSENZAHL, M, 2004b, Beautiful objects as an extension of the self: A reply. *Human–Computer interaction*, **19**, pp. 377–386.

VAN DER HEIJDEN, H., 2003, Factors influencing the usage of websites: The case of a generic portal in the Netherlands. *Information and Management*, **40**, pp. 541–549.

INTERACTIONS, 2004, Special issue on Funology, **11**, September + October.